Assessment

On dimensionality, measurement invariance and suitability of sum scores for the PHQ-9 and the GAD-7

Journal:	*Assessment*
Manuscript ID	ASMNT-20-0125.R1
Manuscript Type:	Original Manuscript
Keywords:	dimensionality, measurement invariance, sum scores, PHQ-9, GAD-7

On dimensionality, measurement invariance and suitability of sum scores for the PHQ-9 and the GAD-7

Jan Stochl, PhD[1,2,3*], Eiko I. Fried, PhD[4], Jessica Fritz, MSc[1], Tim J. Croudace, PhD[5], Debra A. Russo[1], Clare Knight[1], Peter B. Jones, PhD[1,2], Jesus Perez, PhD[1,2]

[1] Department of Psychiatry, University of Cambridge, Cambridge, UK

[2] National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care East of England, Cambridge, UK

[3] Department of Kinanthropology, Charles University, Prague, Czech Republic

[4] Department of Clinical Psychology, Leiden University, Leiden, Netherlands

[5] School of Health Sciences, University of Dundee, Dundee, UK

Corresponding author: Dr Jan Stochl, University of Cambridge, Department of Psychiatry, Herchel Smith Building, Forvie Site, Cambridge Biomedical Campus, Cambridge, CB2 OSZ, UK; Tel: +44 0758 7146299; Fax: +44 1223 336581; Email: js883@cam.ac.uk

Abstract

In psychiatry, severity of mental health conditions and their change over time are usually measured via sum scores of items on psychometric scales. However, inferences from such scores can be biased if psychometric properties such as unidimensionality and temporal measurement invariance for instruments are not met. Here, we aimed to evaluate these properties for common measures of depression (PHQ-9) and anxiety (GAD-7) in a large clinical sample (N=22,362) undergoing psychotherapy. In addition, we tested consistency in dimensionality results across different methods (parallel analysis, factor analysis, explained common variance (ECV), the Partial Credit Model, and the Mokken model). Results show that while both PHQ-9 and GAD-7 are multidimensional instruments with highly correlated factors, there is justification for sum scores as measures of severity. Temporal measurement invariance across 10 therapy sessions was evaluated. Strict temporal measurement invariance was established in both scales, allowing researchers to compare sum scores as severity measures across time.

Keywords: dimensionality, measurement invariance, sum scores, PHQ-9, GAD-7.

Introduction

The assessment of mental health conditions - whether for the purpose of research, screening, diagnostics, or outcome evaluation in therapy - plays a crucial role in psychological and psychiatric research, as well as in clinical practice. Despite progress in recent years, mental health research still lacks biological markers (Prata, Mechelli, & Kapur, 2014; Venkatasubramanian & Keshavan, 2016), and relies largely on questionnaires and scales assessing subjectively rated somatic and psychological symptoms which are hypothesized to be related to candidate diagnostic syndromes (Kapur, Phillips, & Insel, 2012). Therefore, it is of utmost importance that the measurement indicators that are used by clinicians to determine whether someone needs help, benefits from therapy, or progressed to recovery, need to be psychometrically valid and reliable. If not, measurement indicators do not reflect the measured construct and the true progress of the patient. This may lead to patients staying in therapy for an unnecessarily long time, incurring extra cost or being discharged from clinical services before true recovery is reached. This calls for careful assessment of psychometric properties of popular scales.

Both unidimensionality and temporal measurement invariance (hereafter TMI) are critical psychometric properties for scales which are used for assessment of mental health in epidemiological and clinical research as well as in therapeutic practice. Particularly in clinical settings, measurement tools for mental health conditions are often used over time to monitor individual improvement and recovery. Simple sum scores (whether for the total scale or for subscales) are utilized for simplicity and convenience. Unidimensionality is a necessary (yet not sufficient) condition for the meaningful interpretation of sum scores (Heene, Kyngdon, & Sckopke, 2016) and TMI is an additional condition for the meaningful interpretation of sum score changes over time.

Fried and colleagues (2016) investigated unidimensionality and TMI in four common scales for depression (Hamilton Rating Scale for Depression, Quick Inventory of Depressive Symptoms, and two versions of Inventory of Depressive Symptoms (clinical and self-rated)) which routinely use sum scores as a summary statistic in research and clinical practice. They found that both properties did not hold in any of the scales, which challenges "the interpretation of sum scores and their changes as reflecting one underlying construct" (p. 2). Here, our primary aim is to replicate and extend this work by investigating dimensionality and TMI a) for different measurement instruments, b) for depression as well as for anxiety, c) in a larger sample, d) using 10 (rather than 2 time points as in Fried, et al. (2016)), and e) using a more extensive set of methods to explore the issue of uni- vs multidimensionality of the scales from different perspectives.

In this study, we analysed two patient reported outcome measures routinely used to monitor depression and anxiety therapy outcomes in a major UK primary mental health service: the Patient Health Questionnaire-9 (PHQ-9) (Kroenke, Spitzer, & Williams, 2001) and the Generalised Anxiety Disorder assessment-7 (GAD-7) (Spitzer, Kroenke, Williams, & Lowe, 2006). We focused on the following goals:

First, we tested the dimensionality of the item sets comprising the PHQ-9 and GAD-7, treating all responses as ordered categories (ordinal data). In addition, we evaluated whether different psychometric techniques, when applied to the same dataset, provided consistent answers. Dimensionality refers to the number of latent variables that can be estimated from the data and is thus closely related to the scoring of the questionnaire. Indeed, unidimensionality of the instrument (i.e. a single latent variable) is one of the requirements for the justification of using sum scores (the total of the item scores) as summary statistics. This is because, simply put, unidimensionality assures that a single score is a defensible way of scoring each individual (Zwitser & Maris, 2016).

It is, however, not a sufficient condition as it does not say what mathematical form such score should take, i.e. how such a score should be generated. More stringent psychometric requirements may apply to justify using sum scores and they depend on the psychometric model. We discuss sufficient conditions and their evaluation within factor analytic and Item Response Theory frameworks in the Appendix. When an instrument measures multiple constructs, scoring each construct separately (i.e. making sum scores for subscales) may provide more useful and psychometrically sound statistics (Smith, McCarthy, & Zapolski, 2009). However, in both research and clinical practice, sum scores are frequently used without strong empirical evidence for the unidimensionality of the instrument. For example, the Hamilton Rating Scale for Depression (HRSD) (Hamilton, 1960), one of the most commonly used depression measures in clinical practice, is often scored using a sum score of 17 (out of 21) items despite considerable evidence indicating its multidimensionality (Hamilton, 1967; R. Michael Bagby, Andrew G. Ryder, Deborah R. Schuller, & Margarita B. Marshall, 2004; Shafer, 2006). Hamilton himself recommended scoring dimensions separately instead of using a "total crude score" yet these recommendations are regularly ignored. This might also be the case for other questionnaires with a potentially multidimensional structure where the existence of separate constructs are ignored, and unidimensionality is effectively "assumed". In addition, there is sometimes considerable heterogeneity between studies evaluating dimensionality for the same instrument. For example, PHQ-9 and GAD-7 have been investigated by different authors and found to be unidimensional by some (e.g. Gonzalez-Blanch et al., 2018; Lowe et al., 2008) but multidimensional by others (e.g. Beard & Bjorgvinsson, 2014; Elhai et al., 2012).

Second, we test TMI in the PHQ-9 and GAD-7. TMI refers to the degree to which construct validity of the instrument stays stable over time and is thus closely related to the fairness of

temporal comparisons of scores. If TMI holds, changes in the sum score of a given sample represent actual differences in the construct measured through the rating scale (Fried et al., 2016). If TMI does not hold, observed differences in sum scores over time do not necessarily reflect (and cannot be fully attributed to) temporal changes of the latent variable. We provide a TMI investigation, comparing PHQ-9 and GAD-7 across 10 timepoints.

Apart from extending the work of Fried, et al. (2016), this study has three additional aims. The first one is to investigate whether various methods for dimensionality assessment provide consistent outcomes when the results of their analyses are compared. The second one is to argue and showcase that multidimensional scales may still be usefully summarized using a sumscore. The third one is to illustrate a number of different psychometric techniques that can be used for the assessment of dimensionality. We provide statistical code to implement each method and synthetic data. We hope this will enable readers to adopt our examples, explore these methods, and conduct sets of evaluations on their own data.

Methods

Setting

The Improving Access to Psychological Therapies (IAPT) programme in England began in 2008 with a direct objective to improve access to evidence-based psychological treatment for common mental disorders (CMD) such as anxiety and depression. The programme has continued to expand over time and currently assesses over 1,6 million people with CMD annually, delivering therapy to approximately 1.06 million people. It aims to increase public access to psychological therapies approved by the National Institute for Health and Care Excellence (NICE) through offering

flexible referral routes (including self-referral and stepped care pathways). Accordingly, the IAPT programme provides low (step 2) or high-intensity (step 3) treatment to people aged 16+ years. Low-intensity IAPT approaches include guided self-help, psychoeducation, computerised CBT, behavioural activation, and structured group physical activity programmes.(Clark, 2018) In high-intensity IAPT services, face-to-face cognitive–behavioural therapy is the predominant approach, although there is a wider range of recommended treatments (e.g. eye movement desensitization and reprocessing (EMDR), interpersonal psychotherapy (IPT), counselling for depression, compassion-focused therapy (CFT), and Integrative Counselling). In high-intensity IAPT, patients receive 7 sessions on average over a period of 3-4 months. Nationally, recovery rates exceed 52%, about quarter of patients (25.7%) do not improve, and small percentage (5.8%) deteriorate. Drop-out rates are relatively high (appr. 46%).

Primary measures: PHQ-9 and GAD-7

At each therapy session, IAPT therapists routinely assess depression and anxiety symptomatology using the 9-item PHQ-9 (Kroenke et al., 2001) and the 7-item GAD-7 questionnaire(Spitzer et al., 2006). Both scales were adopted by the IAPT programme nationally because of their sound validity (Cameron, Crawford, Lawton, & Reid, 2008; Maroufizadeh, Omani-Samani, Almasi-Hashiani, Amini, & Sepidarkish, 2019; Spitzer et al., 2006; Titov et al., 2011), reliability (Johnson, Ulvenes, Øktedalen, & Hoffart, 2019b; Maroufizadeh et al., 2019), sensitivity and specificity (Levis, Benedetti, & Thombs, 2019; Spitzer et al., 2006) and brevity. They are officially used to monitor recovery rates across all IAPT services. Total scores on both instruments are computed as a sum score of items (response categories are identical for both instruments: 0=Not at all; 1=Several days,

2=More than half the days; 3=Nearly every day). Thus, PHQ-9 scores can range from 0 to 27, where cores of 5, 10, 15, and 20 represent cutpoints for mild, moderate, moderately severe and severe depression, respectively. GAD-7 scores can range between 0 and 21. Scores of 5, 10, and 15 represent cut-off points for mild, moderate, and severe anxiety, respectively. In IAPT, individuals are described as at 'caseness', if they score above the clinical cut-off for depression (PHQ-9 $\geq$ 10) (Manea, Gilbody, & McMillan, 2012) *or* anxiety (GAD-7 $\geq$ 8) and are in recovery if they score below these thresholds for *both* measures. Here, we have analysed the PHQ-9 and GAD-7 data from the first 10 therapy appointments.

Participants

We included all IAPT patients from two trusts (Cambridge and Peterborough Foundation Trust and Sussex Partnership NHS Foundation Trust) who received services between February and December 2018. Data from 22,362 individuals was available for the first therapy session of which 66.4% were women, 33.3% were men, and 0.3% had missing data on gender. Mean age of the sample was 40.1 years (sd=15.4 years). Most individuals in the sample were white (88.2%) and the remainder was divided into four ethnicity categories (1.1% were Indian, 0.8% asian, 0.7% black, and 2.4% stated mixed or other ethnicity background). Information on ethnicity for 6.8% of patients was missing. An average patient severity at the start of the therapy was moderate, with sum score mean of 13.6 for PHQ-9 (sd=6.28) and 12.6 (sd=5.3) for GAD-7. Histograms of sum scores for both measures are provided in Supplementary Figure S1 and S2.

The sample size decreased considerably as available therapy session data increased, due to both dropout and discharge of patients. Yet, a subsample of 6,554 individuals had PHQ-9 and GAD-7

scores for 10 therapy sessions. Sample sizes, means and standard deviations for PHQ-9 and GAD-7 total scores for each therapy appointment are available in Figure 1.

Statistical Analysis

For the assessment of dimensionality we examined the number of factors needed to describe each questionnaire at each therapy session. A large number of psychometric approaches were used to test dimensionality including a) parallel analysis (Horn, 1965), b) exploratory factor analysis (EFA), c) confirmatory factor analysis (CFA), d) parametric (Rasch) item response theory (IRT) model, e) nonparametric IRT (Mokken) model, and f) explained common variance (ECV). It is important to note that for the sake of brevity and clarity, we only report outcomes of analyses relevant for dimensionality assessment. Thus, some typical or recommended outcomes of these psychometric techniques are missing. This note is specifically relevant for Partial Credit Model and the Mokken model.

Confirmatory factor analysis (CFA). We first assessed the fit of a 1-factor model at each measurement point (therapy appointment) to evaluate whether unidimensionality can be justified using a confirmatory approach. The CFA model fit was considered good if the root mean square error of approximation (RMSEA) was 0.06 or lower, standardised root mean squared residual (SRMR) was 0.08 or lower, and the comparative fit index (CFI) was 0.95 or higher (Hu & Bentler, 1999). We have considered that items are ordinal and used mean and variance adjusted weighted least squares (WLSMV) as the estimator. We used *MPlus* software (L. K. Muthén & Muthén, 1998-2019) to estimate CFA models.

Parallel analysis (PA) and exploratory factor analysis (EFA). In the case that unidimensional models using CFA did not fit the data, we used parallel analysis to determine the number of factors. In order to compare results with (Fried et al., 2016) we mimicked their setting for parallel analysis. To this end, we compared the observed eigenvalues with eigenvalues of randomly drawn data, and we extracted factors for which the eigenvalues exceeded the randomly generated eigenvalues (50 parallel datasets for each analysis and used 95% eigenvalue percentiles). We used the function *fa.parallel* from the R-package *psych* (Revelle, 2018). Using EFA (in *MPlus*) with a WLSMV estimator, we have assessed the fit of models with 2-5 factors (note that 1-factor model was tested using CFA) with oblimin factor rotations. The most parsimonious model which met the same fit criteria as described above for the CFA model was then selected.

Partial Credit Model (PCM). A PCM (Masters, 1982) is a model for polytomous item responses from a family of Rasch models and therefore shares the distinguishing characteristics of that family: separable person and item parameters, raw scores as sufficient statistics (i.e. the sum score carries all the information about the measured attribute of the respondent), and, hence, conjoint item score additivity (Masters & Wright, 1997). A good fit of data with Rasch model provides stringent support for the existence of a single, quantitative and unidimensional psychological variable underlying the scale items (Glas & Verhelst, 1995; Heene et al., 2016). We therefore conclude unidimensionality when all items fit the PCM. Fit is evaluated using indices such as outfit and infit. These statistics are based on standardized residuals, which are the standardized differences between the observations and their expected values according to the Rasch model. Their sum of squares approximates a χ^2 distribution and the outfit is simply the ratio of the χ^2 and its degrees of freedom (Wright & Masters, 1990). Infit is an information-weighted form of outfit. The weighting reduces the influence of less informative, low variance, off-target responses. The

expected value of outfit and infit is 1.0 and ranges from 0 to infinity. Values larger than 1.0 indicate unmodeled noise on a ratio scale (e.g. 1.1 indicates 10% excess noise). Values less than 1.0 indicate overfit of the data to the model, i.e., too predictable observations (Linacre, 2002). Rating scales items (such as those of the PHQ-9 and GAD-7 have an acceptable fit when these indices range between 0.6 and 1.4 (Wright & Linacre, 1994). For this analysis we used R package *eRm* (Mair & Hatzinger, 2007).

Mokken model. We also empirically assessed the questionnaire dimensionality within the framework of Mokken models (Mokken, 1971) using R package *mokken* (van der Ark, 2012). Mokken models are often seen as a non-parametric version of Rasch models (Stochl, Jones, & Croudace, 2012). For this we used Loevingers' item scalability coefficients cutoffs (Loevinger, 1947), which we according to recommendations increased from 0.3 up to 0.45 (in 0.05 increments) (Stochl et al., 2012). Note that we did not aim to evaluate other constituting properties of the Mokken models (monotonicity and non-intersection of item response functions, local independence) but we simply used this approach as an automated engine to explore how it would build unidimensional (sub)scales of the instrument (Gillespie, Tenvergert, & Kingma, 1987; van der Ark, 2012). Unidimensionality was concluded if the engine extracted only a single Mokken scale and, at the same time, all items from the corresponding instruments were included in this scale.

Omega hierarchical **(ω_H)** ***and explained common variance (ECV).*** Hierarchical omega (ω_H) is the coefficient proposed by McDonald (1999) which estimates the proportion of variance in total scores that can be attributed to a single general factor. Hierarchical omega can also be interpreted as the reliability coefficient (the larger the coefficient, the more accurately one can predict an individual's relative standing on the latent variable common to all the scale's indicators based on

their observed scale score) and as the generalizability coefficient (square of the correlation between the scale score and the latent variable common to all the indicators) (Revelle, 2018). To calculate ω_H we used a function in the R-package *psych* (Revelle, 2018) which estimates a factor model with oblique factor rotation and performs the Schmid Leiman transformation to find general factor loadings and then calculates the index itself. The explained common variance (ECV) is an index similar to ω_H in terms of interpretation, but superior to ω_H as an index of unidimensionality as it utilises only the reliable variance of the sum scores (P. M. Bentler, 2009; Reise, Moore, & Haviland, 2010; Ten Berge & Sočan, 2004). ECV was computed based on formula provided by Reise et al. (2010). Both ω_H and ECV were used to evaluate the extent to which scores reflect a single latent variable even when the data are multidimensional, i.e. in the presence of more than one highly related sub-dimensions. Hence, even if the questionnaires are multidimensional, sum scores may be justified, if the percentage of ECV is high.

Temporal measurement invariance (TMI). The assessment of TMI was conducted as an iterative process during which we increased equality constraints on the most parsimonious well-fitting factor structure for both instruments obtained from EFA, correspondingly testing configural (M1), weak (M2), strong (M3), or strict (M4) invariance. As a first step, a configural invariance model M1 was fit to the data of all measurement points per instrument; the model imposes no equality constraints on the parameters, and only restricts the number of factors to be equal across time. In the next step, the weak factorial invariance model M2 was estimated; M2 constrains item loadings to be equal across time. The strong factorial invariance model M3 additionally constrains thresholds to be equal across time, and the strict invariance model M4 forces all residual invariances to be equal on top of all previous constraints. Once estimated, each model is compared to the previous one with respect to the fit to the data. If introducing equality constraints decreases

the fit significantly, measurement invariance is rejected. TMI can be established only if M4 is not rejected (Meredith, 1993). We refer the reader to B. Muthén and Asparouhov (2013) for a thorough descriptions of these constraints within MPlus, and Millsap (2011) for a general overview and interpretation of TMI models.

Code availability

To help the reader conduct our analyses on their own data, we provide the analysis code at https://osf.io/r2e63/.

Data availability

Data were made available for analysis as part of an exploratory evaluation project (forming part of an NIHR programme grant for applied research number PG-0616-20003); due to the confidentiality and protection of the original dataset we were not allowed to provide the data. However, we created synthetic data with almost identical descriptive statistics, distributional properties and covariances/correlations using R package *synthpop* (Nowok, Raab, & Dibben, 2016). The synthetic data can be used to mimic the analyses carried out in this paper and is available online at https://osf.io/r2e63/.

Results

Description of PHQ-9 and GAD-7 sum scores by cumulative appointments

We show means and standard deviations for the PHQ-9 and the GAD-7 sum scores in Figure 1. Those scores suggest that patients improve over time in both depression and anxiety, and the heterogeneity of the sum scores is similar across appointments (the variances appear not to vary). Distribution of sumscores is depicted in Supplementary Figures S1 and S2.

---------- insert Figure 1 about here ----------

Assessment of Dimensionality

Confirmatory factor analysis

Fit indices for unidimensional models for PHQ-9 and GAD-7 across therapy sessions are reported in Table 1. For both instruments, goodness-of-fit of the 1-factor model varied per fit index and provided a somewhat conflicting message. The CFI index which compares the 1-factor model with estimated factor loadings and factor variance constrained to 1 to the null model (i.e. the model where all factor loadings equal 1 and variance of the factor is set to zero) showed an acceptable fit regardless of the time point. Similarly, the SRMR, the fit index evaluating the size of residual correlations, showed good fit across time points. On the other hand, RMSEA values showed a consistently poor fit for both the PHQ-9 and the GAD-7. There is no clear explanation of inconsistency between RMSEA and other indices as it may stem from the nonlinear interplay between fit of the baseline model and degrees of freedom of the model (Lai & Green, 2016).

---------- insert Table 1 about here ----------

Parallel Analysis and Exploratory Factor Analysis

Parallel analysis (PA) suggested that both instruments have a multidimensional structure, although one dominant factor emerged for both instruments at all time points. For the PHQ-9, four factors were extracted with exception of the 9th appointment for which 3 factors described the data best. For the GAD-7, two factors were extracted for 8 out of 10 timepoints and three factors were identified at appointment 1 and 7.

The EFA analyses showed consistent results across time. The minimal number of factors to achieve good fit (i.e. model having CFI over 0.95 and, at the same time, RMSEA below 0.06) was 3 for the PHQ-9 and 2 for GAD-7. The factorial structure (outlined as note under the Table 2) was stable across time for both instruments. These findings are presented in Table 2.

---------- insert Table 2 about here ----------

Partial Credit Model

The item fit for the partial credit model (PCM) is presented in Table 3. Both infit and outfit were in the range for an acceptable item fit across all time points (0.6-1.4). This indicates that all items fit the PCM, which supports a unidimensional factorial structure for both scales.

---------- insert Table 3 about here ----------

Mokken model

Table 4 shows abridged results of fitting a Mokken model across therapy appointments. For both instruments, a single Mokken scale was extracted based on recommended Loevingers' item scalability coefficient (H_i) threshold of 0.3 (Loevinger, 1947; Mokken, 1971). No items were excluded. We gradually increased the cutoff in line with recommendations up to 0.45 (Stochl et al., 2012), but the results did not change. This provides empirical justification for the unidimensionality of both instruments. In addition, the scalability coefficient H (a measure of strength of the extracted unidimensional scale) was over 0.5 (with exception of session 1 for PHQ-9 where H=0.482) which is indicative of "strong homogeneity/unidimensionality" of the extracted scale (Sijtsma & Molenaar, 2002; Stochl et al., 2012).

---------- insert Table 4 about here ----------

Hierarchical omega and explained common variance

Based on the hierarchical omega and ECV values in Figure 2, we can conclude that across appointments, 79%-86% of the variance (73%-80% of reliable variance) of the sum score of PHQ-9 and 76%-85% of the variance (74%-84% of reliable variance) of the sum score of GAD-7 is attributable to variance on the corresponding general factor. Interpretations of ω_H allow for two additional conclusions: a) reliability of both instruments is satisfactory and b) correlation between

sum score and the corresponding general latent variable lies between 0.89 and 0.93 for PHQ-9 and between 0.87 and 0.92 for GAD-7 (computed as square roots of the ω_H).

---------- insert Figure 2 about here ----------

Assessment of temporal measurement invariance (TMI)

Fit indices of models with constraints specific to each level of TMI are presented in Table 5. Note that TMI constraints are imposed on the most parsimonious well-fitting factor structure derived from the EFA models (3-factor for PHQ-9 and 2-factor for GAD-7). Results are similar for both instruments. Chi-square values suggest significant difference across TMI models, but this finding is expected in large samples regardless of true model differences. All other fit indices suggest negligible differences in fit between configural, weak, strong and strict invariance models. The fact that the strict invariance models do not fit worse compared to corresponding configural models supports the notion that TMI holds for both the PHQ-9 and the GAD-7. Interestingly, RMSEA and CFI show marginal superiority for more constraint models.

---------- insert Table 5 about here ----------

Discussion

Recently, the concern has been raised that measurement of depression over time is problematic due to violations of psychometric properties that permit usage of sum scores as suitable summary

statistics (Fried, 2017; Fried et al., 2016; Shafer, 2006). Such concern is particularly relevant to mental health research as well as clinical practice in which sum scores are often used to monitor change of both depression and anxiety over time. This study aimed to investigate the dimensionality and TMI for two widely used depression and anxiety scales routinely used to monitor therapy outcomes in primary mental health services in the UK.

Dimensionality

Three of the five applied approaches (PA, EFA and CFA) suggested a multidimensional structure of both scales. Parametric (Partial Credit Model) and nonparametric (Mokken model) IRT approaches, however, supported a unidimensional structure. These results do not need to be seen as conflicting. In our interpretation of the models, there is evidence for multidimensionality in both scales, but these dimensions are highly correlated. The ECV and hierarchical omega coefficients, which were derived from bifactor model framework, suggested that the structure of both scales is dominated by a strong general factor capturing around 80% of the variance of all items. Therefore, we argue that the main finding supports the use of sum scores as a suitable summary statistic for both the PHQ-9 and the GAD-7.

In the literature, factor structures reported for these instruments are inconsistent. For PHQ-9, previous studies reported unidimensional (Gonzalez-Blanch et al., 2018; Keum, Miller, & Inkelas, 2018) as well as 2-dimensional structures (Elhai et al., 2012; Guo et al., 2017; Chilcot et al., 2013; Krause, Reed, & McArdle, 2010; Richardson & Richards, 2008), consisting of somatic and affective factors. Reported GAD-7 structures include unidimensional (Lowe et al., 2008; Sousa et al., 2015), modified unidimensional (Bartolo, Monteiro, & Pereira, 2017; Johnson, Ulvenes,

Øktedalen, & Hoffart, 2019a; Lee & Kim, 2019), or 2-factors (Beard & Bjorgvinsson, 2014; Kertz, Bigda-Peyton, & Bjorgvinsson, 2013). We believe that this inconsistency may stem from the methodological plurality in the literature, where different methods support different conclusions—very similar to our own investigation.

Temporal Measurement Invariance (TMI)

We compared the fit of increasingly constrained models to evaluate temporal measurement invariance (TMI) of PHQ-9 and GAD-7. For the configural model we used the most parsimonious well-fitting factor structure derived from the EFA models (3-factors for PHQ-9 and 2-factors for GAD-7). The fit was similar across configural, weak, strong, and strict invariance models. Chi-square differences between models were found, but ignored due to our extremely large sample size, in which case the use of this statistic is not recommended (P.M. Bentler, 1990). CFI, TLI and RMSEA indices even showed a slightly superior fit for more constrained models. These results suggest that measurement invariance holds and provide empirical justification for the comparability of scores across time (Cheung & Rensvold, 2002).

Previously, TMI was supported for a two-dimensional PHQ-9 solution (Elhai et al., 2012; Guo et al., 2017). The results of previous studies where PHQ-9 was considered as unidimensional measure, are both positive (Gonzalez-Blanch et al., 2018) and negative (Downey, Hayduk, Curtis, & Engelberg, 2016) with regard to TMI. Studies for GAD-7 are scarce but homogeneous in support of TMI (Mewton, Hobbs, Sunderland, Newby, & Andrews, 2014; Naragon-Gainey, Gallagher, & Brown, 2014).

Strengths and Limitations

This study benefits from a primary care sample that is not only large, but is also fairly representative of the clinical population seeking psychological therapies (Knight et al., 2020, in preparation). However, the average number of therapy sessions was 8 in our sample, but 7 in the general IAPT sample. This may indicate that our sample is a little less treatment responsive than the 'general' IAPT population.

Our study has several limitations. First, the sample has a notable attrition due to dropout from therapy or discharge of individuals when they reach recovery; only about 30% of the original sample seen at baseline had 10 or more appointments. This is expected because the average number of appointments in IAPT is 7 (NHS digital, 2020). In our sample, 55.5% completed scheduled treatment, 22.11% of cases dropped out before their treatment was finished (the end of care reason was unknown for 15.3% of cases and the remaining cases were discontinued for various reasons, for example, discharge to secondary care). Arguably, the sub-sample of individuals with a large number of appointments is structurally different from the original sample as it consists of individuals who need/require more treatment. As such, we do not necessarily see such structural differences as a limitation. For example, the fact that the dimensionality and factorial structure are the same across appointments (and thus potentially across structurally/qualitatively different subsamples) may indicate measurement invariance across classes of individuals who respond differentially to IAPT therapy. We suggest that conjectures regarding sub-group invariance should further be evaluated in future studies.

Second, a potential constraint may be that the patients were allocated to therapies of different intensity: less severe cases are allocated to low intensity therapy (46.6% of our sample) and more complex/severe cases (53.4%) into high intensity therapy; this was not taken into account in our

analyses. Therefore, we cannot be sure that we would have revealed unidimensional and TMI had we tested the models separately per treatment arm. On the other hand, this can also be seen as an advantage because it suggests that the unidimensionality and the TMI of PHQ-9 and the GAD-7 hold up in a natural setting taking various different treatment interventions together into a single sample.

Third, we did not evaluate the meaningfulness of sum scores nor the validity of the studied scales from a content validity perspective. Indeed, the item coverage of the PHQ-9 and the GAD-7 may not be ideal. Thus, although the measures seem to be fairly unidimensional and invariant, they may not evaluate the disorders in their full breadth. However, this limitation is not specific to the measures scrutinized here, and applies across mental health measures (Fried, 2017).

Fourth, as indicated above, temporal invariance does not imply sub-group measurement invariance, which we decided to not investigate in this study. In other words, even if PHQ-9 and GAD-7 scores may adequately reflect within-individual changes of the disorder, such scores may not provide fair comparison across sub-groups such as gender or ethnicity.

Finally, a technical limitation is that MPlus does not provide robust maximum likelihood estimation to estimate likelihood-based fit indices such as the Akaike Information Criterion and Bayesian Information Criterion (BIC) which would provide a more straightforward comparison of TMI models.

Conclusion

Our results show that both PHQ-9 and GAD-7 can be considered as multidimensional measures but with a strong corresponding general factor which explains around 80% of the variance of

unweighted sum score of items. Hence, we propose that using sum scores for either scale seems is acceptable. In addition, temporal measurement invariance appears to hold for both scales. This supports the conjecture that meaningful comparisons of sum scores of the PHQ-9 and the GAD-7 over time are justified, which is crucial for longitudinal research as well as for monitoring outcomes in clinical practice.

Author Contributions

JS, JF and EF designed analysis plan and wrote the first draft, JS carried out data analysis and interpretation and had full access to all the data in the study. All other authors contributed to subsequent versions of the manuscript. JS takes responsibility for the integrity of the data and the accuracy of the data analysis.

Acknowledgements

We are extremely grateful to the IAPT teams who participated in this study and provided access to the required data. In addition, we thank the Norwich Clinical Trials Unit (NCTU) for its support managing exports of data.

Conflicts of Interest Disclosures

JS discloses consultancy for IESO digital health. The remaining authors have no conflicts of interest.

Funding

This paper presents independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (Reference Number RP-PG-0616-20003). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. PBJ, JP and JS received support from the NIHR Applied Research Collaboration (ARC) East of England (NIHR200177). JF is funded by the Medical Research Council Doctoral Training/Sackler Fund and the Pinsent Darwin Fund.

References

Bartolo, A., Monteiro, S., & Pereira, A. (2017). Factor structure and construct validity of the Generalized Anxiety Disorder 7-item (GAD-7) among Portuguese college students. *Cad Saude Publica, 33*(9), e00212716. doi:10.1590/0102-311X00212716

Beard, C., & Bjorgvinsson, T. (2014). Beyond generalized anxiety disorder: psychometric properties of the GAD-7 in a heterogeneous psychiatric sample. *J Anxiety Disord, 28*(6), 547-552. doi:10.1016/j.janxdis.2014.06.002

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bentler, P. M. (2009). Alpha, Dimension-Free, and Model-Based Internal Consistency Reliability. *Psychometrika, 74*(1), 137-143. doi:10.1007/s11336-008-9100-1

Cameron, I. M., Crawford, J. R., Lawton, K., & Reid, I. C. (2008). Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. *Br J Gen Pract, 58*(546), 32-36. doi:10.3399/bjgp08X263794

Clark, D. M. (2018). Realizing the Mass Public Benefit of Evidence-Based Psychological Therapies: The IAPT Program. *Annu Rev Clin Psychol, 14*, 159-183. doi:10.1146/annurev-clinpsy-050817-084833

Downey, L., Hayduk, L. A., Curtis, J. R., & Engelberg, R. A. (2016). Measuring Depression-Severity in Critically Ill Patients' Families with the Patient Health Questionnaire (PHQ): Tests for Unidimensionality and Longitudinal Measurement Invariance, with Implications for CONSORT. *J Pain Symptom Manage, 51*(5), 938-946. doi:10.1016/j.jpainsymman.2015.12.303

Elhai, J. D., Contractor, A. A., Tamburrino, M., Fine, T. H., Prescott, M. R., Shirley, E., . . . Calabrese, J. R. (2012). The factor structure of major depression symptoms: A test of four competing models using the Patient Health Questionnaire-9. *Psychiatry Research, 199*(3), 169-173. doi:https://doi.org/10.1016/j.psychres.2012.05.018

Fried, E. I. (2017). Moving forward: how depression heterogeneity hinders progress in treatment and research. *Expert Review of Neurotherapeutics, 17*(5), 423-425. doi:10.1080/14737175.2017.1307737

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time . . . Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment, 28*(11), 1354-1367. doi:10.1037/pas0000275

Gillespie, M., Tenvergert, E. M., & Kingma, J. (1987). Using Mokken scale analysis to develop unidimensional scales. *Quality and Quantity, 21*(4), 393-408. doi:10.1007/bf00172565

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. Fischer & I. W. Molenaar (Eds.), *Rasch Models - Foundations, Recent Developments, and Applications* (pp. 69-95). New York, NY:: Springer.

Gonzalez-Blanch, C., Medrano, L. A., Munoz-Navarro, R., Ruiz-Rodriguez, P., Moriana, J. A., Limonero, J. T., . . . Psic, A. P. R. G. (2018). Factor structure and measurement invariance across various demographic groups and over time for the PHQ-9 in primary care patients in Spain. *PLoS One, 13*(2), e0193356. doi:10.1371/journal.pone.0193356

Guo, B., Kaylor-Hughes, C., Garland, A., Nixon, N., Sweeney, T., Simpson, S., . . . Morriss, R. (2017). Factor structure and longitudinal measurement invariance of PHQ-9 for specialist

mental health care patients with persistent major depressive disorder: Exploratory Structural Equation Modelling. *J Affect Disord, 219*, 1-8. doi:10.1016/j.jad.2017.05.020

Hamilton, M. (1960). A rating scale for depression. *J Neurol Neurosurg Psychiatry, 23*, 56-62. doi:10.1136/jnnp.23.1.56

Hamilton, M. (1967). Development of a rating scale for primary depressive illness. *The British journal of social and clinical psychology, 6*(4), 278-296. doi:10.1111/j.2044-8260.1967.tb00530.x

Heene, M., Kyngdon, A., & Sckopke, P. (2016). Detecting Violations of Unidimensionality by Order-Restricted Inference Methods. *Frontiers in Applied Mathematics and Statistics, 2*(3). doi:10.3389/fams.2016.00003

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185. doi:10.1007/bf02289447

Hu, L., & Bentler, M. P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233-255. doi:10.1207/S15328007SEM0902_5

Chilcot, J., Rayner, L., Lee, W., Price, A., Goodwin, L., Monroe, B., . . . Hotopf, M. (2013). The factor structure of the PHQ-9 in palliative care. *J Psychosom Res, 75*(1), 60-64. doi:10.1016/j.jpsychores.2012.12.012

Johnson, S. U., Ulvenes, P. G., Øktedalen, T., & Hoffart, A. (2019a). Psychometric Properties of the General Anxiety Disorder 7-Item (GAD-7) Scale in a Heterogeneous Psychiatric Sample. *Frontiers in Psychology, 10*(1713). doi:10.3389/fpsyg.2019.01713

Johnson, S. U., Ulvenes, P. G., Øktedalen, T., & Hoffart, A. (2019b). Psychometric Properties of the General Anxiety Disorder 7-Item (GAD-7) Scale in a Heterogeneous Psychiatric Sample. *Frontiers in Psychology, 10*, 1713-1713. doi:10.3389/fpsyg.2019.01713

Kapur, S., Phillips, A. G., & Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol Psychiatry, 17*(12), 1174-1179. doi:10.1038/mp.2012.105

Kertz, S., Bigda-Peyton, J., & Bjorgvinsson, T. (2013). Validity of the Generalized Anxiety Disorder-7 scale in an acute psychiatric sample. *Clin Psychol Psychother, 20*(5), 456-464. doi:10.1002/cpp.1802

Keum, B. T., Miller, M. J., & Inkelas, K. K. (2018). Testing the factor structure and measurement invariance of the PHQ-9 across racially diverse U.S. college students. *Psychological Assessment, 30*(8), 1096-1106. doi:10.1037/pas0000550

Knight, C., Russo, D., Stochl, J., Croudace, T., Fowler, D., Grey, N., . . . Perez, J. (2020). Prevalence of and Recovery from Common Mental Disorder including Psychotic Experiences in the UK Primary Care Improving Access to Psychological Therapies (IAPT) Programme. *Journal of Affective Disorders, In preparation*.

Krause, J. S., Reed, K. S., & McArdle, J. J. (2010). Factor Structure and Predictive Validity of Somatic and Nonsomatic Symptoms From the Patient Health Questionnaire-9: A Longitudinal Study After Spinal Cord Injury. *Archives of Physical Medicine and Rehabilitation, 91*(8), 1218-1224. doi:https://doi.org/10.1016/j.apmr.2010.04.015

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9 - Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*(9), 606-613. doi:DOI 10.1046/j.1525-1497.2001.016009606.x

Lai, K., & Green, S. B. (2016). The Problem with Having Two Watches: Assessment of Fit When RMSEA and CFI Disagree. *Multivariate Behavioral Research, 51*(2-3), 220-239. doi:10.1080/00273171.2015.1134306

Lee, B., & Kim, Y. E. (2019). The psychometric properties of the Generalized Anxiety Disorder scale (GAD-7) among Korean university students. *Psychiatry and Clinical Psychopharmacology, 29*(4), 864-871. doi:10.1080/24750573.2019.1691320

Levis, B., Benedetti, A., & Thombs, B. D. (2019). Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ, 365*, l1476. doi:10.1136/bmj.l1476

Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs, 61*(4).

Lowe, B., Decker, O., Muller, S., Brahler, E., Schellberg, D., Herzog, W., & Herzberg, P. Y. (2008). Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Medical care, 46*(3), 266-274. doi:10.1097/MLR.0b013e318160d093

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9), 1-20.

Manea, L., Gilbody, S., & McMillan, D. (2012). Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *CMAJ, 184*(3), E191-196. doi:10.1503/cmaj.110829

Maroufizadeh, S., Omani-Samani, R., Almasi-Hashiani, A., Amini, P., & Sepidarkish, M. (2019). The reliability and validity of the Patient Health Questionnaire-9 (PHQ-9) and PHQ-2 in patients with infertility. *Reproductive health, 16*(1), 137-137. doi:10.1186/s12978-019-0802-x

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Masters, G. N., & Wright, B. D. (1997). The Partial Credit Model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 101-121). New York, NY: Springer New York.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah: Lawrence Erlbaum Associates, Inc.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543.

Mewton, L., Hobbs, M. J., Sunderland, M., Newby, J., & Andrews, G. (2014). Reductions in the internalising construct following internet-delivered treatment for anxiety and depression in primary care. *Behaviour research and therapy, 63*, 132-138. doi:10.1016/j.brat.2014.10.001

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Taylor and Francis Group.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton.

Muthén, B., & Asparouhov, T. (2013). Version 7.1 MPlus Language Addendum. Retrieved from https://www.statmodel.com/download/Version7.1xLanguage.pdf

Muthén, L. K., & Muthén, B. O. (1998-2019). *Mplus User's Guide. Eighth Edition*. Los Angeles, CA: Muthén & Muthén.

Naragon-Gainey, K., Gallagher, M. W., & Brown, T. A. (2014). A longitudinal examination of psychosocial impairment across the anxiety disorders. *Psychological Medicine, 44*(8), 1691-1700. doi:10.1017/S0033291713001967

NHS digital. (2020). Psychological Therapies: reports on the use of IAPT services, England October 2019 Final including reports on the IAPT pilots. Retrieved from https://digital.nhs.uk/data-and-information/publications/statistical/psychological-therapies-report-on-the-use-of-iapt-services/october-2019-final-including-reports-on-the-iapt-pilots

Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software; Vol 1, Issue 11 (2016)*.

Prata, D., Mechelli, A., & Kapur, S. (2014). Clinically meaningful biomarkers for psychosis: a systematic and quantitative review. *Neurosci Biobehav Rev, 45*, 134-141. doi:10.1016/j.neubiorev.2014.05.010

R. Michael Bagby, Ph.D. ,, Andrew G. Ryder, M.A. ,, Deborah R. Schuller, M.D. , and, & Margarita B. Marshall, B.Sc. (2004). The Hamilton Depression Rating Scale: Has the Gold Standard Become a Lead Weight? *American Journal of Psychiatry, 161*(12), 2163-2177. doi:10.1176/appi.ajp.161.12.2163

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. *J Pers Assess, 92*(6), 544-559. doi:10.1080/00223891.2010.496477

Revelle, W. (2018). psych: Procedures for Personality and Psychological Research (Version 1.8.12). Northwestern University, Evanston, Illinois, USA. Retrieved from https://CRAN.R-project.org/package=psych

Richardson, E. J., & Richards, J. S. (2008). Factor structure of the PHQ-9 screen for depression across time since injury among persons with spinal cord injury. *Rehabilitation Psychology, 53*(2), 243.

Shafer, A. B. (2006). Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of Clinical Psychology, 62*(1), 123-146. doi:10.1002/jclp.20213

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). London: Sage Publications.

Smith, G. T., McCarthy, D. M., & Zapolski, T. C. (2009). On the value of homogeneous constructs for construct validation, theory testing, and the description of psychopathology. *Psychological Assessment, 21*(3), 272-284. doi:10.1037/a0016699

Sousa, T. V., Viveiros, V., Chai, M. V., Vicente, F. L., Jesus, G., Carnot, M. J., . . . Ferreira, P. L. (2015). Reliability and validity of the Portuguese version of the Generalized Anxiety Disorder (GAD-7) scale. *Health and Quality of Life Outcomes, 13*(1), 50. doi:10.1186/s12955-015-0244-2

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Lowe, B. (2006). A brief measure for assessing generalized anxiety disorder - The GAD-7. *Archives of Internal Medicine, 166*(10), 1092-1097. doi:DOI 10.1001/archinte.166.10.1092

Stochl, J., Jones, P. B., & Croudace, T. J. (2012). Mokken scale analysis of mental health and wellbeing questionnaire item responses: a nonparametric IRT method in empirical

research for applied health researchers. *BMC Med Res Methodol, 12*(1), 74. doi:10.1186/1471-2288-12-74

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69*(4), 613-625. doi:10.1007/bf02289858

Titov, N., Dear, B. F., McMillan, D., Anderson, T., Zou, J., & Sunderland, M. (2011). Psychometric comparison of the PHQ-9 and BDI-II for measuring response during treatment of depression. *Cogn Behav Ther, 40*(2), 126-136. doi:10.1080/16506073.2010.550059

van der Ark, L. A. (2012). New Developments in Mokken Scale Analysis in R. *2012, 48*(5), 27. doi:10.18637/jss.v048.i05

Venkatasubramanian, G., & Keshavan, M. S. (2016). Biomarkers in Psychiatry - A Critique. *Annals of neurosciences, 23*(1), 3-5. doi:10.1159/000443549

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright, B. D., & Masters, G. N. (1990). Computation of OUTFIT and INFIT statistics. *Rasch Measurement Transactions, 3*(4), 84-85.

Figure Captions

Figure 1: Means and standard deviations for PHQ-9 and GAD-7 sum scores across 10 therapy appointments

Figure 2: Omega hierarchical and estimated common variance across 10 therapy appointments

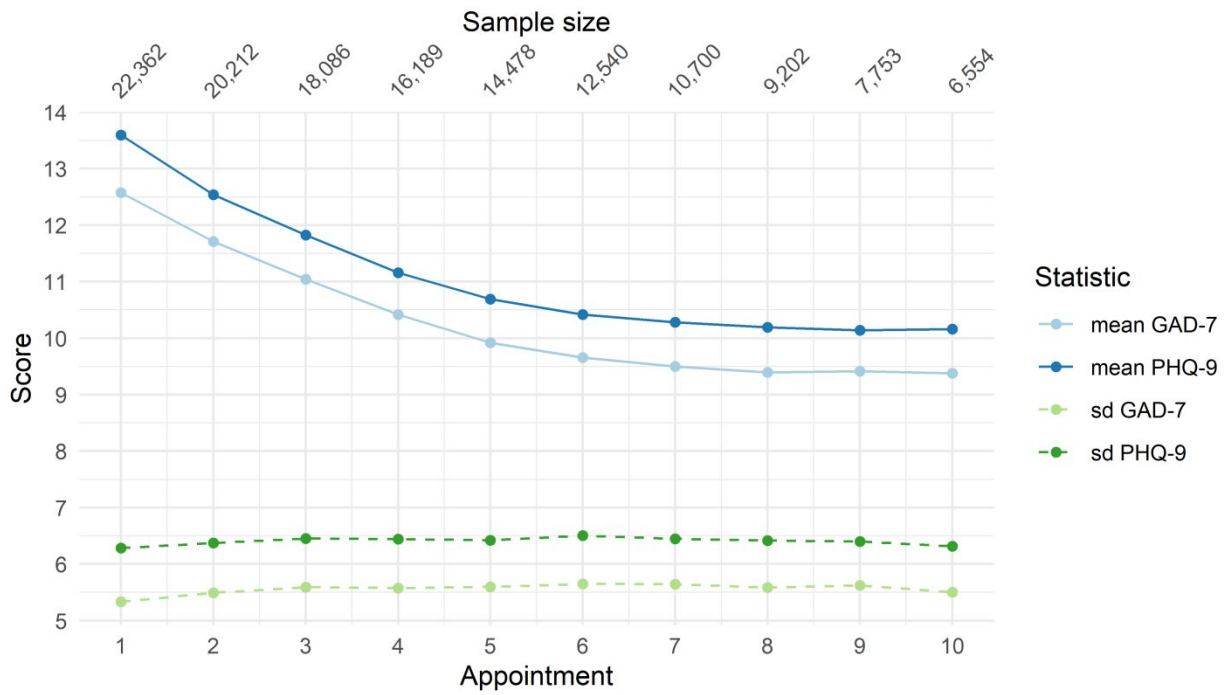

Figure 1: Means and standard deviations for PHQ-9 and GAD-7 sum scores across 10 therapy appointments

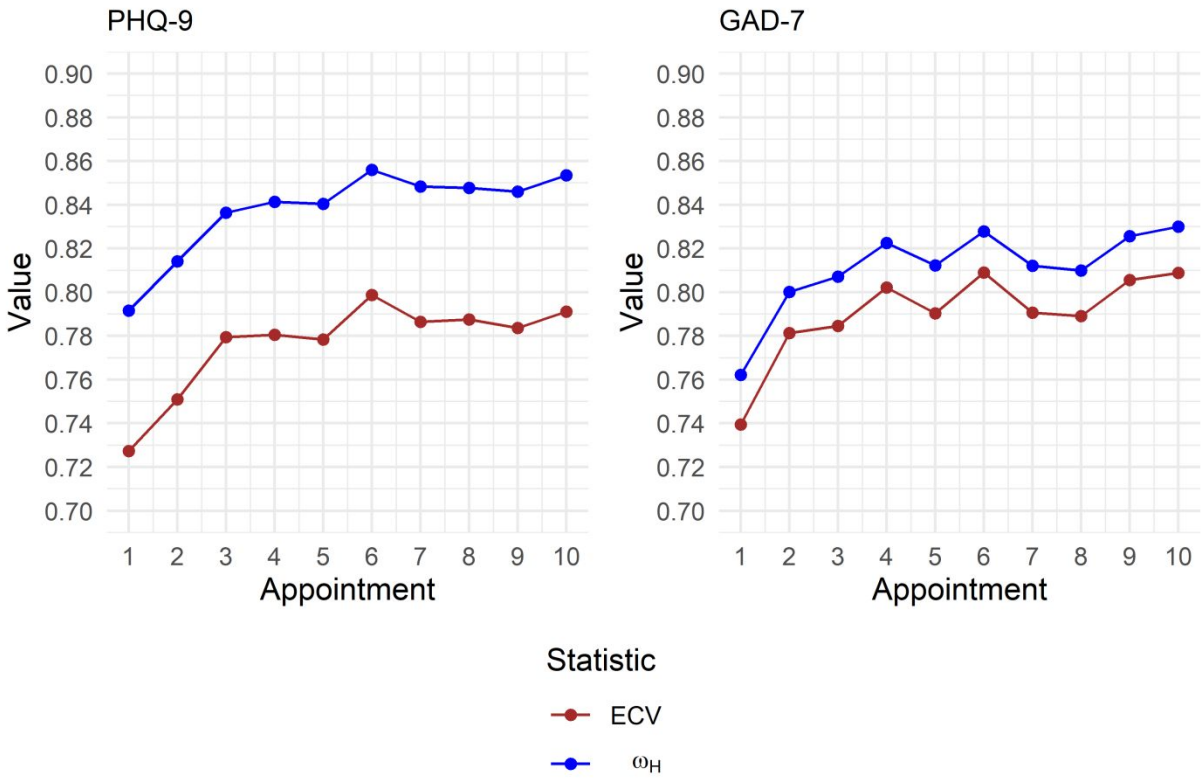

Figure 2: Omega hierarchical and estimated common variance across 10 therapy appointments

Table 1: Fit indices of CFA

Appointment nr.	PHQ-9				GAD-7			
	$\chi 2$ (27)	CFI	RMSEA (90% CI)	SRMR	$\chi 2$ (14)	CFI	RMSEA (90% CI)	SRMR
1	8,013*	0.957	0.115 (0.113,0.117)	0.045	6,723*	0.977	0.146 (0.143,0.149)	0.042
2	7,602*	0.963	0.118 (0.115,0.12)	0.043	2,337*	0.982	0.159 (0.154,0.165)	0.037
3	6,900*	0.967	0.119 (0.116,0.121)	0.041	1,777*	0.986	0.148 (0.142,0.153)	0.034
4	6,858*	0.967	0.125 (0.122,0.127)	0.042	1,663*	0.986	0.152 (0.146,0.158)	0.036
5	6,326*	0.967	0.127 (0.124,0.129)	0.042	1,439*	0.986	0.153 (0.146,0.159)	0.035
6	5,283*	0.971	0.124 (0.122,0.127)	0.04	1,202*	0.985	0.151 (0.144,0.158)	0.034
7	4,778*	0.969	0.128 (0.125,0.131)	0.042	944*	0.986	0.144 (0.136,0.152)	0.033
8	4,161*	0.969	0.129 (0.125,0.132)	0.042	850*	0.984	0.15 (0.142,0.159)	0.034
9	3,652*	0.968	0.131 (0.128,0.135)	0.042	709*	0.983	0.151 (0.142,0.16)	0.035
10	2,979*	0.968	0.129 (0.125,0.133)	0.042	520*	0.986	0.141 (0.131,0.152)	0.032

* $p<0.001$

Table 2: Fit indices of the EFA models which satisfy close fit (CFI>0.95 and RMSEA<0.06) across 10 therapy appointments

Appointment nr.	PHQ-9				GAD-7			
	nr. Factors PA	nr. Factors EFA[a]	CFI	RMSEA	nr. Factors PA	nr. Factors EFA[b]	CFI	RMSEA
1	4	3	0.996	0.051	3	2	0.998	0.051
2	4	3	0.997	0.049	2	2	0.999	0.049
3	4	3	0.998	0.049	2	2	0.999	0.045
4	4	3	0.997	0.051	2	2	0.999	0.046
5	4	3	0.998	0.048	2	2	0.999	0.05
6	4	3	0.998	0.046	2	2	0.999	0.05
7	4	3	0.998	0.053	3	2	0.999	0.051
8	4	3	0.998	0.046	2	2	0.999	0.053
9	3	3	0.998	0.043	2	2	0.999	0.047
10	4	3	0.998	0.047	2	2	0.999	0.049

[a] Factor 1: 'Interest', 'Hopeless', 'Feeling Bad', 'Hurt'; factor 2: 'Asleep', 'Tired', 'Appetite'; factor 3: 'Concentrate', 'Moving'. Mean (sd) factor correlations: factor 1 and 2=0.799 (0.034); factor 1 and 3=0.819 (0.029); factor 2 and 3=0.778 (0.045).

[b] Factor 1: 'Nervous', 'Cannot Control Worry', 'Worry Too Much', 'Afraid'; factor 2: 'Trouble Relax', 'Restless', 'Annoyed'. Mean (sd) factor correlation=0.731 (0.027).

Please see Supplementary Table S1 for full item wording.

Table 3: Item Fit indices of the PCM models across 10 therapy appointments

Appointment nr.	PHQ-9		GAD-7	
	range Outfit	range Infit	range Outfit	range Infit
1	0.68 - 1.12	0.70 - 1.09	0.62 - 1.32	0.64 - 1.28
2	0.68 - 1.12	0.69 - 1.09	0.61 - 1.30	0.62 - 1.25
3	0.68 - 1.12	0.69 - 1.11	0.61 - 1.28	0.62 - 1.25
4	0.69 - 1.12	0.69 - 1.11	0.60 - 1.30	0.62 - 1.25
5	0.69 - 1.10	0.70 - 1.09	0.61 - 1.26	0.62 - 1.23
6	0.68 - 1.14	0.68 - 1.10	0.61 - 1.25	0.62 - 1.23
7	0.68 - 1.11	0.69 - 1.09	0.59 - 1.28	0.60 - 1.25
8	0.69 - 1.09	0.70 - 1.13	0.62 - 1.25	0.62 - 1.22
9	0.68 - 1.11	0.69 - 1.10	0.60 - 1.26	0.60 - 1.23
10	0.69 - 1.11	0.69 - 1.17	0.62 - 1.28	0.63 - 1.25

Table 4: Results of Mokken Automatic Item Selection Procedure across 10 therapy appointments. Note that there was always single Mokken scale found and no items were excluded.

Appointment nr.	PHQ-9		GAD-7	
	H (SE)	range Hi	H (SE)	range Hi
1	0.482 (0.003)	0.420 - 0.547	0.549 (0.003)	0.440 - 0.619
2	0.515 (0.003)	0.454 - 0.578	0.587 (0.003)	0.492 - 0.651
3	0.546 (0.003)	0.481 - 0.604	0.613 (0.003)	0.525 - 0.674
4	0.562 (0.004)	0.497 - 0.617	0.622 (0.004)	0.530 - 0.683
5	0.572 (0.004)	0.516 - 0.624	0.634 (0.004)	0.548 - 0.694
6	0.590 (0.004)	0.535 - 0.645	0.649 (0.004)	0.562 - 0.703
7	0.588 (0.004)	0.540 - 0.642	0.651 (0.004)	0.561 - 0.711
8	0.587 (0.005)	0.520 - 0.640	0.649 (0.005)	0.565 - 0.704
9	0.588 (0.005)	0.533 - 0.645	0.652 (0.005)	0.565 - 0.710
10	0.584 (0.006)	0.507 - 0.638	0.643 (0.005)	0.548 - 0.698

Notes: H = scale scalability coefficient; SE=standard error; Hi = item scalability coefficient

Table 5: Temporal measurement invariance across 10 therapy appointments

	Model	N	nr. of parameters	χ^2 diff (df)	χ^2 diff p-value	CFI	TLI	RMSEA (90% CI)	SRMR
PHQ-9	Configural (M1)	23,631	876	-	-	0.937	0.926	0.037 (0.037,0.038)	0.057
	Weak (M2)	23,631	759	1,429 (117)	<0.001	0.938	0.929	0.036 (0.036,0.037)	0.057
	Strong (M3)	23,631	678	2,479 (198)	<0.001	0.937	0.930	0.036 (0.036,0.036)	0.057
	Strict (M4)	23,631	607	2,362 (269)	<0.001	0.945	0.939	0.034 (0.033,0.034)	0.057
GAD-7	Configural (M1)	23,610	533	-	-	0.951	0.944	0.042 (0.042,0.043)	0.052
	Weak (M2)	23,610	436	1,502 (97)	<0.001	0.951	0.946	0.041 (0.041,0.042)	0.052
	Strong (M3)	23,610	373	2,396 (160)	<0.001	0.951	0.947	0.041 (0.041,0.041)	0.052
	Strict (M4)	23,610	318	2,426 (215)	<0.001	0.956	0.954	0.038 (0.038,0.039)	0.053

Appendix: Notes on additional conditions for sufficiency and interpretability of sum scores

In the manuscript, we refer to dimensionality and temporal measurement invariance (TMI) as necessary yet not sufficient conditions for meaningful interpretation of sum scores as summaries of the latent variable intended to be measured. Unidimensionality assures that the measured latent variable can be summarized using a single score for each person (Zwitser & Maris, 2016). TMI condition assures that the internal structure of the measure remains the same across measurement occasions – in that case the meaning of the summary score does not change and the differences of scores are interpretable as differences in the measured latent variable. Therefore, TMI is not essential for cross-sectional applications.

However, neither unidimensionality nor TMI inform what mathematical form the summary score should take. For example, does simple sum of raw responses to scale items accurately represents the measured latent variable or should we weight raw responses because items vary in psychometric quality? To answer such question, more stringent psychometric requirements may need to be applied. Here, we aim to a) briefly discuss additional analytical criteria for considering the sum score as a sufficient and interpretable statistic for the ordering of individuals, and b) provide links between terminology used across two psychometric frameworks - factor analysis and Item Response Theory (IRT). Apart from analytical aspects discussed here, additional conceptual requirements and validity checks may apply to justify sum scores, but these are not discussed here.

Simple sum score is inherently unweighted, or sometimes called unit-weighted score (i.e. each item has the same weight, typically equal to 1). Implicitly, this means that each item in the scale a) has the same importance, b) contributes an equal amount of information to the construct being measured, and c) is equally valid. For unidimensional models, it also means all items have the same correlation with the latent variable. This translates into constraints that

need to be introduced for the parameters of the psychometric models. In factor analytic framework, this constraint is that items in the factor model need to be *parallel*, i.e. have equal loadings and equal error variances. Indeed, correlation between sum scores and factor scores from a parallel model equal 1 (McNeish & Wolf, 2020) and thus sum scores are a perfect linear transformation of factor scores in this model. When item loadings and measurement errors vary across items (known as the *congeneric* model), the weighted sum score is a more suitable approach to scoring. The weights assigned to each item reflect psychometric quality of items and enable those with higher loadings (and thus smaller measurement error) to contribute more. This means that the higher is the variability of factor loadings the less appropriate is the use of (unweighted) sum score. An excellent discussion on the sum score as a constrained form of factor analytic model is provided in McNeish and Wolf (2020).

In the IRT framework, the 'slope' of the item response curves (conditional probabilities of each categorical response option as a function of measured latent variable) is referred to as item discrimination. It informs how well the item differentiates between those with high and low levels of the measured latent variable. Loadings in categorical data factor analysis are straightforwardly related to item discriminations through simple formulas (e.g. McDonald, 1999, p. 259). Therefore, the constraint of equal loadings in factor analysis is equivalent to that of equal IRT item discriminations. IRT models with equal item discrimination are considered as Rasch type models (Rasch, 1960). When these models fit the data well, then the unweighted sum score is a sufficient statistic for ordering individuals with respect to the measured latent variable.

Note that individuals with identical sum score can have very different item response patterns (Fried & Nesse, 2015). In this sense, sum score is "item free" meaning it does not matter *which* items are endorsed. In IRT literature, this is referred to as a feature of Rasch model

called “specific objectivity” (comparisons between individuals become independent of which particular items have been used, see Rasch (1977)).

In order to offer a simple visual representation, let’s consider a 5-item depression measure with categorical response options that are binary scored (symptom absent=0, symptom present=1). In this case, equal discriminations mean that item response curves will all be parallel to each other, as in the left panel of Figure A1 (note the similarity with parallel model in factor analysis).

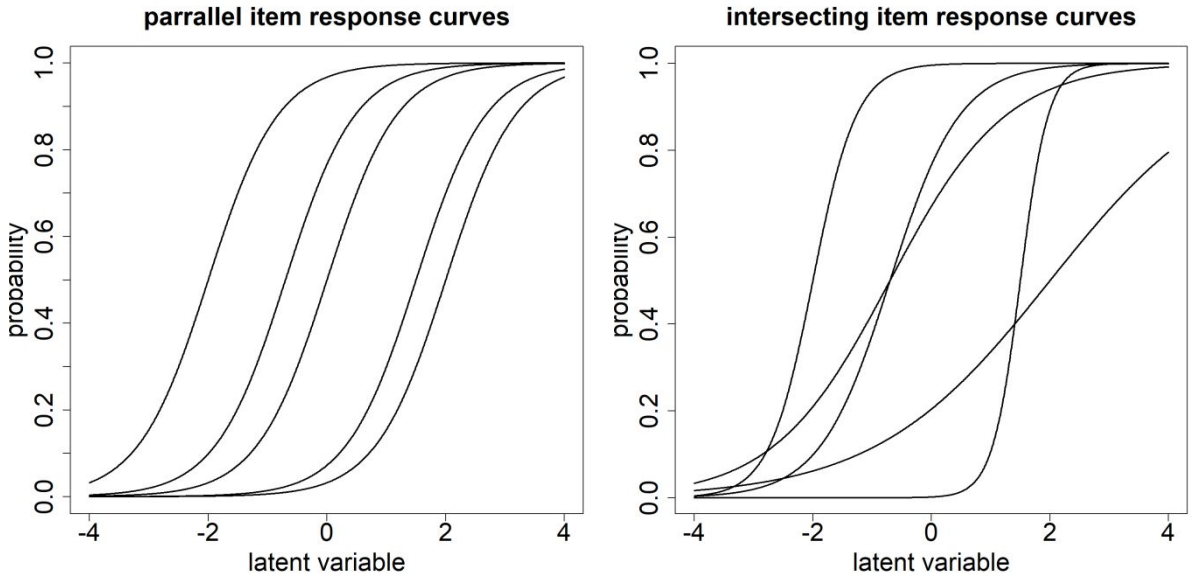

Figure A1: Examples of item response curves for parallel (left) and non-parallel (right) for binary scored items

IRT models where discriminations are freely estimated across items and thus these models are conceptually equivalent to congeneric model in factor analysis include two-parameter logistic model (2-PLM) for dichotomously scored items and General partial credit model (Muraki, 1992) or graded response model (Samejima, 1969) for polytomously scored items. For 2-

PLM, this is displayed in the right panel of Figure A1 where item response curves are no longer parallel and may (yet not necessarily) intersect. For 2-PLM model, the weighted sum score is sufficient statistic for ordering individuals with respect to the measured latent variable (Zwitser & Maris, 2016).

We showed that unweighted sum scores are justified when factor loadings, or equivalently discriminations, are equal across items. However, a weaker condition referred to as the monotone likelihood ratio (MLR) has also been also proposed in the literature as a sufficient condition to order subjects based on their unweighted sum scores (Hemker, Sijtsma, Molenaar, & Junker, 1997). This encouraged applied researchers to use models from within the nonparametric IRT framework, such as Mokken models (Mokken, 1971) for the purpose of sum score justification. However, importantly Zwitser and Maris (2016) showed that for dichotomous items MLR justifies using sum scores for *group* comparisons but is not in fact a sufficient condition for the ordering of *individuals*. Instead, these authors proposed another sufficient condition denoted as "ordinal sufficiency". They showed, that in the class of IRT models suited for dichotomous items this condition is met only by the Rasch model (and in the nonparametric Rasch model they introduce).

As shown above, the sum score is formally also representable as a highly constrained unidimensional factor analytic model or in various related ways by unidimensional models from the IRT family. Constrained unidimensional models can be tested against less constrained ones and thus suitability of sum scores can be statistically tested. For example, testing difference in fit measures between parallel and congeneric unidimensional factor analytic models, or similarly between Rasch and 2-PLM can provide evidence whether unit-weighting sum score is reasonable.

References

Fried, E. I., & Nesse, R. M. (2015). Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR* D study. *Journal of Affective Disorders, 172*, 96-102.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62*(3), 331-347.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah: Lawrence Erlbaum Associates, Inc.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*. doi:10.3758/s13428-020-01398-0

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Rasch, G. (1977). On Specific Objectivity: An Attempt of Formalizing the Generality and Validity of Scientific Statements. *Danish Yearbook of Philosophy, 14*, 58-94.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph no 17.*

Zwitser, R. J., & Maris, G. (2016). Ordering Individuals with Sum Scores: The Introduction of the Nonparametric Rasch Model. *Psychometrika, 81*(1), 39-59. doi:10.1007/s11336-015-9481-x

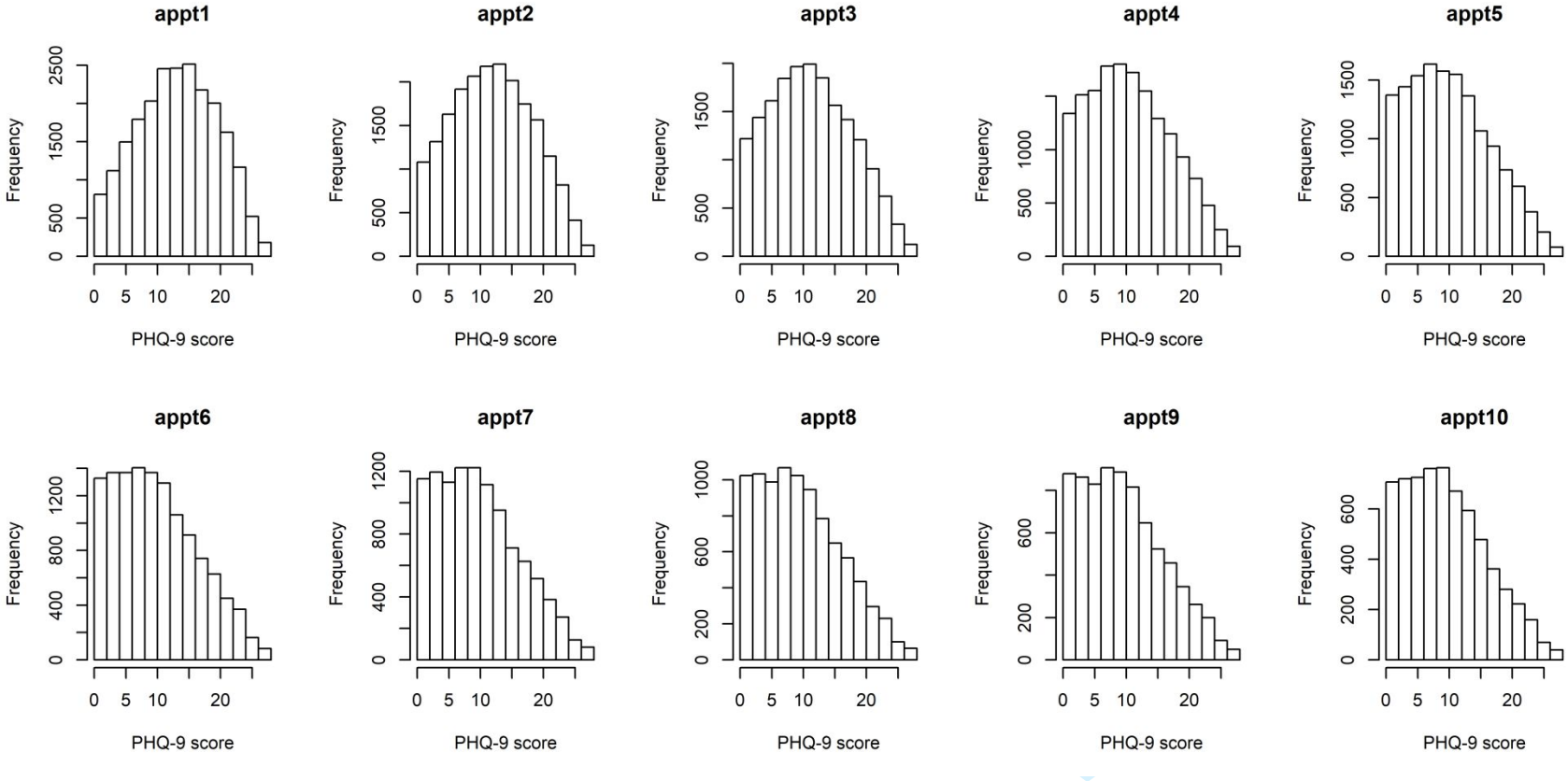

Figure S1: Histogram of sum scores across 10 therapy appointments for PHQ-9

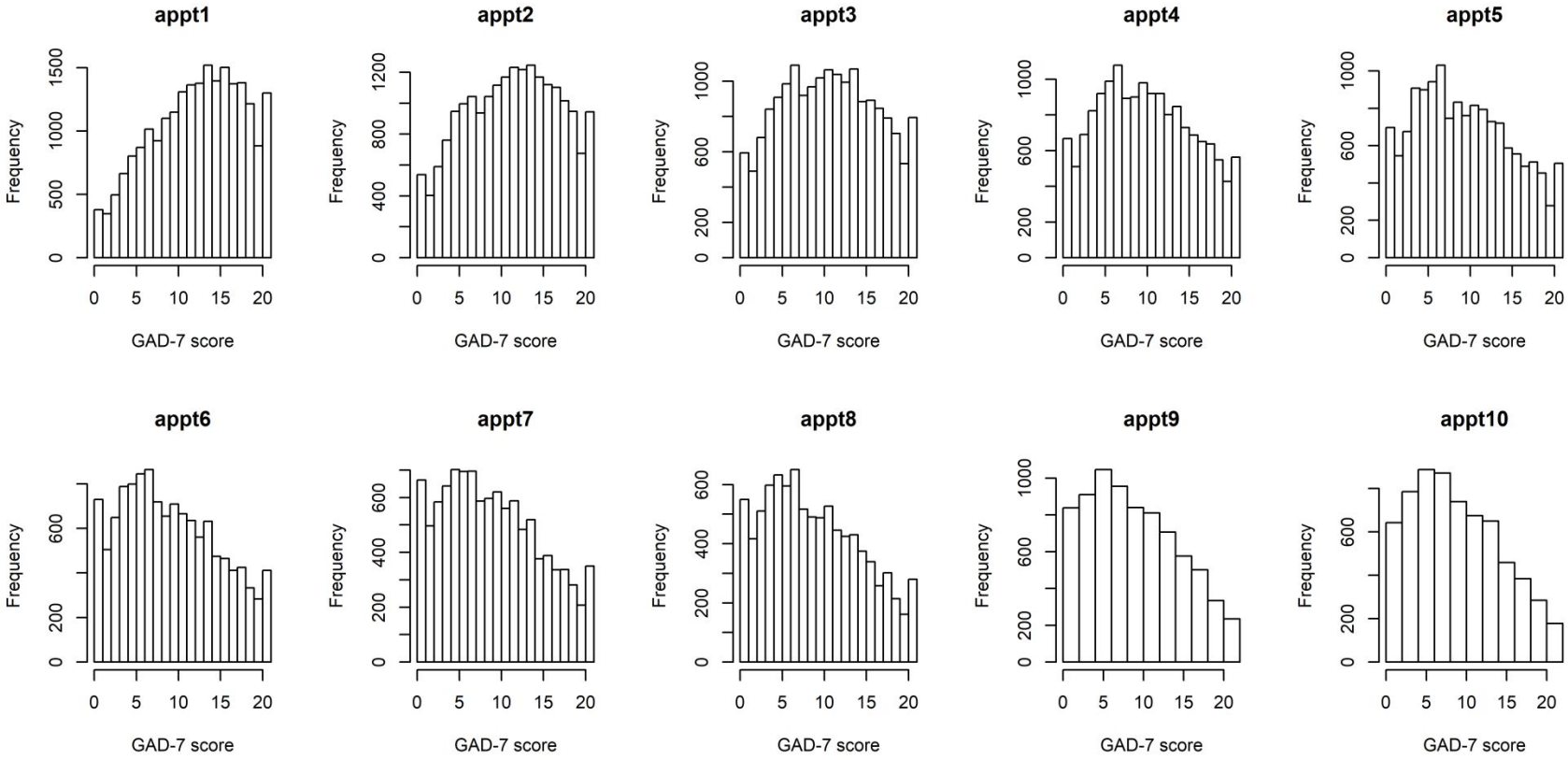

Figure S2: Histogram of sum scores across 10 therapy appointments for GAD-7

Table S1: PHQ-9 and GAD-7 item labels and wording

Measure	Item	Statement
PHQ-9	Interest	Little interest or pleasure in doing things
	Hopeless	Feeling down, depressed, or hopeless
	Asleep	Trouble falling/staying asleep, sleeping too much
	Tired	Feeling tired or having little energy
	Appetite	Poor appetite or overeating
	Feeling Bad	Feeling bad about yourself or that you are a failure or have let yourself or your family down
	Concentrate	Trouble concentrating on things, such as reading the newspaper or watching television
	Moving	Moving or speaking so slowly that other people could have noticed. Or the opposite; being so fidgety or restless that you have been moving around a lot more than usual
	Hurt	Thoughts that you would be better off dead or of hurting yourself in some way.
GAD-7	Nervous	Feeling nervous, anxious or on edge
	Cannot Control Worry	Not being able to stop or control worrying
	Worry Too Much	Worrying too much about different things
	Trouble Relax	Trouble relaxing
	Restless	Being so restless that it is hard to sit still
	Annoyed	Becoming easily annoyed or irritable
	Afraid	Feeling afraid as if something awful might happen